Szótárazási dilemmák a MetaMorpho magyar-angol fordítóprogram névszói adatbázisának építésében

Vincze Veronika[1], Lucza Mónika[1], Csendes Dóra[1], Kiss Gabriella[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Nyelvtechnológiai Csoport
H-6720 Szeged, Árpád tér 2.
{vinczev, lucza, dcsendes}@inf.u-szeged.hu
[2] MorphoLogic Kft. Budapest
H-1126 Budapest, Orbánhegyi út 5.
gkiss@morphologic.hu

Kivonat: Jelen cikk a MetaMorpho magyar-angol fordítóprogram kétnyelvű szótárának előállítása során tapasztalt gyakorlati nehézségekről és azok nyelvészeti hátteréről számol be. A szótár fejlesztésében a Szegedi Tudományegyetem a magyar névszói kifejezések – azaz főnévi kifejezések (NX), melléknévi (ADJX) és határozószói szótári bejegyzések (ADVX) – angolra fordításával vette ki részét. Az így létrejövő szótári adatbázis jelenleg közel 90 ezer bejegyzést tartalmaz. A névszói kifejezések fordítása során a célkitűzés az volt, hogy a gyakorisági elemzések alapján a magyar nyelvhasználatnak leginkább megfelelő jelentés kerüljön első helyre az adatbázisban, ugyanakkor igyekeztünk a lehető legtöbb jelentést felvenni.

1 Bevezetés

A gépi fordítórendszerek teljesítményét nagyban befolyásolja a rendszerben lévő szótári adatbázis minősége. A szótárak jelentőségét növeli még az a tény is, hogy a felhasználók a szótár bővítésével interaktív módon javítani tudják a rendszert. A létező fordítórendszerek szótárai a lexikai elemeket igen eltérő formátumban, tartalmi lefedettséggel, részletességgel, és eltérő precizitású formalizmussal tartalmazzák [2]. A szótári adatbázisok megvalósítása mindig az adott rendszer sajátosságainak függvénye. Az interlingva rendszerek szótárainak például természetükből fakadóan nem szükséges fordítási információkat tartalmazniuk, míg a sok nyelvre készülő fordítóprogramok esetében gyakori, hogy az alkalmazott nyelvekre részletes egynyelvű szótárakkal is rendelkeznek a kétnyelvű, a transzfer során használt szótárak mellett [1].

A MetaMorpho fordítóprogram-család a különféle fordítási feladatokhoz kíván eszközöket biztosítani [5,6,7]. A rendszer elemzési szabályokon alapszik, amelyekhez fordítások vannak közvetlenül hozzárendelve. Ezek a szabályok kódolják a különböző nyelvtani szabályszerűségeket, lexikális elemeket, vonzatkereteket, szemantikai jegyeket és egyéb mintákat.

A jelen tanulmány a MetaMorpho magyar-angol fordítóprogram kétnyelvű szótári adatbázisának előállítása során tapasztalt gyakorlati nehézségekről és azok nyelvészeti hátteréről számol be. A szótári adatbázis kialakítása két fő fázisban történt: az automatikus előfeldolgozási, illetve előfordítási fázist egy manuális ellenőrzési és javítási fázis követte. Az előfeldolgozás során sor került a kifejezések gépi felhasználásra alkalmas egységes szabály formátumra hozására, valamint potenciális fordítási javaslatok automatikus módon történő előállítására. Ezt követően a kézi ellenőrzés során arra törekedtünk, hogy minél szélesebb körű információval lássunk el egy-egy bejegyzést, így azok nem csupán a kifejezések fordítását tartalmazzák, hanem morfoszintaktikai, illetve szemantikai információkat is. Igyekeztünk továbbá figyelmet fordítani arra is, hogy jellemzően csak az alapszókincsbe tartozó szavak kerüljenek be az adatbázisba, így az elavult, tájnyelvi vagy szakzsargonba tartozó szavakat kiszűrtük.

2 Automatikus előfeldolgozás

A szótári adatbázis szócikkei három különböző forrásból származnak: (1) a MetaMorpho angol-magyar fordítóprogram anyagának kifordításával és feldolgozásával keletkezett szótári egységek; (2) a MorphoLogic által korpusz-gyakorisági vizsgálat alapján létrehozott szótári egységek; (3) a Nyelvtudományi Intézet és a Szegedi Tudományegyetem által létrehozott és jegyekkel ellátott szótári egységek. Az előfeldolgozás célja a fordítandó nyelvi forrás összegyűjtése, mmd-formátumra[68] [3] hozása, ill. emberi fordítási folyamat lerövidítése volt.

Az (1) csoport szótári egységeinek létrehozása volt a legnagyobb körültekintéssel végzendő feladat, mivel itt az angol-magyar szótári adatbázis visszafordításból származó azonos magyar fordításokat kellett egységesíteni, a nekik megfelelő angol fordításokat pedig alternatív megoldásként felsorolni. Ezekben az esetekben az angol-magyar fordítóprogramban szereplő szemantikai jegyeknek a kifordított szabályba való automatikus átemelése sokszor nem volt egyértelmű. Bizonyos szemantikai jegyek (pl. *földrajzi nevek, tulajdonneveket leíró szemantikai jegyek*) átvétele viszont kevésbé volt problémás.

A (2) csoport szótári egységeinek a feldolgozását a Magyar Nemzeti Szövegtár[69] alapján végeztük el. Egy Perl-script segítségével határoztuk meg azt, hogy az alapszótár alapján javasolt fordítás vagy fordítások milyen arányban fordulnak elő a Hunglish[70] kétnyelvű korpusz angol részeiben. Ennek alapján a lehetséges fordításokra egy sorrendet állítottunk fel és a legvalószínűbb fordítással kitöltöttük a szabály angol fordítását.

A (3) csoport szótári egységei egynyelvűek voltak, a szótári kifejezések el voltak látva szemantikai jegyekkel is. Itt a kitöltetlen generáló oldal (angol oldal) automatikus feltöltése az (2) csoportban ismertetett automatikus módszerrel zajlott.

Az automatikusan előállított szabályokat átnéző nyelvészek feladata az volt, hogy a fordítás helyességét ellenőrizzék és megjegyzésben jelöljék észrevételeiket a sza-

[68] Az mmd-formátum (lásd 3. fejezet példái) a MetaMorpho keretrendszer nyelvi adatainak és szabályainak formátuma.

[69] http://corpus.nytud.hu/mnsz/

[70] http://szotar.mokk.bme.hu/hunglish/search/corpus

bályra vonatkozóan. A szabályokhoz tartozó megjegyzéseket osztályokba soroltuk, ezekhez egységes jelölésrendszert alkottunk, amely alapján szükség szerint pl. adott szabályok egységesen kivonhatók lettek a rendszerből. Az (1), (2) és a (3)-mal jelzett kifejezésekhez tartozó szabályok százalékos aránya sorrendben a jelenlegi rendszerben a következőképpen alakult *45,3:34,4: 20,3:%*. Jelenleg összesen 73.795 főnévi, 12.449 melléknévi, valamint 3270 határozói kifejezést tartalmaz a szótári adatbázis. A fordítórendszer hatékonyságának növelése érdekében ezen felül közel 4500 kollokáció is szerepel az adatbázisban.

A magyar-angol források bővítése természetesen nem zárult le, később a szótárt a felhasználók visszajelzései alapján tovább bővítjük. A MorphoLogicban fejlesztés alatt van egy terminológia-kivonatoló, amivel a fordítóprogram felhasználója interaktív módon módosíthatja, ill. bővítheti a szótári adatbázist.

3 A fordításszabályok felépítése

Az automatikus előkészítés eredményeképpen a következő, ún. mmd-formátumú szabályok álltak elő. Egy mmd-szabály kötelezően áll egy fejlécből, egy elemző sorból és egy vagy több generáló sorból. Az mmd-szabály tartalmazhat megjegyzés sorokat is. Az mmd-szabály elemző ill. generáló sora funkcionálisan két részre bontható: feltétel- és értékadást leíró részre. A feltétel mindig kerek zárójelek között van megfogalmazva, a feltételt leíró elemek formailag jegy és érték párok, ahol az értékek sztring és szimbólum típusúak lehetnek. A szabály értékadás része kapcsos zárójelek között található. Az értékadás formai leírása is jegy-érték párokon alapul, annak a használata a feltételek elemeivel megegyező.

Példa:
```
*ADJX=alkoholmentes:304
HU.ADJX = ADJ(lex="alkoholmentes")
EN.ADJX = N[lex="alcohol"] + PUNCT[lex="gluehyphen"] + ADJ[lex="free"]
;cmt: adj.mmd
;lexs: |non gluehyphen alcoholic,alcohol gluehyphen free
;tr_A: non-alcoholic
```

A szabály első sorából kiolvasható az adott kifejezés nyelvtani kategóriája (jelen példában ez ADJX-szel jelölt melléknév). A második sor a magyar (elemzési oldal), a harmadik sor az angol (generálási oldal) kifejezést adja meg részeire bontva. A magyar oldalon levő melléknévnek az angol oldalon egy főnév és egy melléknév kötőjellel összekapcsolt kombinációja felel meg. A szabály tartalmazza még az előfeldolgozás során előállított fordítási javaslatokat is (jelölésük: ;tr_A). A javasolt fordítások azonban nem mindig helytállóak, így ezekben az esetekben az ellenőrzést végzőknek kellett keresni megfelelő fordítást az adott kifejezésre. A szabályokban továbbá szerepelhetnek még morfológiai, szemantikai, szórendre, illetve szóhasználatra vonatkozó információk is. Jelölni lehet például, ha a magyar és angol kifejezés száma eltér, illetve a főnevek esetét is feltüntetjük, amennyiben nem alanyesetben állnak (morfológiai információ):

HU.NX[ennum=PL] = N(lex="tutyi")
EN.NX = N#1[lex="carpet"] + N[lex="slipper"]

HU.NX = N(lex="tehéntej")
EN.NX = N#1[lex="cow", case=GEN] + N[lex="milk"]

A melléknévi kifejezések fordításakor megjelöljük, ha a melléknév angol megfelelője – az általános szabálytól eltérően – a főnév mögött helyezkedik el (szórendi információ):

HU.ADJX[enpos=POST] = ADJ(lex="mustáros")
EN.ADJX(:head="PREP") = PREP[lex="with"] + N[lex="mustard"]

Utóbbi szabály egyben a szintaktikai fej jelölésére is példa. A fejet csak akkor jelöljük, ha nem esik egybe a névszói csoport alapértelmezett fejével

4 A névszói csoportok fordítása során szerzett tapasztalatok

A MetaMorpho fordítóprogram szótári adatbázisának fejlesztésében a Szegedi Tudományegyetem a magyar névszói kifejezések – azaz főnévi és melléknévi kifejezések – és határozószói szótári bejegyzések angolra fordításával vette ki részét. A fejlesztés során nem csupán az egyes bejegyzések angol megfelelőjének megadása volt a feladat, hanem szükség esetén további nyelvészeti, használatbeli jellegzetességekre vonatkozó információt is tárolni lehetett a szabályokban. Fontosnak bizonyult, hogy a gyakorisági vizsgálatok alapján a magyar nyelvhasználatnak leginkább megfelelő jelentés kerüljön első helyre a szótárban, mivel a jelenlegi rendszer egy jelentést kezel. A későbbi fejlesztésekre való tekintettel lehetőség szerint igyekeztünk egy adott bejegyzés minél több jelentését felvenni.

4.1 A főnévi csoportok fordítása során szerzett tapasztalatok

A főnévi kifejezések fordítása kapcsán felmerült problémákat három fő típusba lehet sorolni. Az első problémakörbe a többjelentésű szavak tartoznak, hiszen bizonyos esetekben nehezen lehetett eldönteni, hogy az adott bejegyzésnek melyik jelentése a (leg)gyakoribb (pl.: *ráhajtás*, *partnercsere*, *fazon*). A második csoportba azok a szavak tartoznak, ahol a szótári információk nem teljesen helytállóak, például hibás fordítás szerepel a magyar–angol szótárban (*vattapamacs*, *szülőszék*). A harmadik tipikus probléma a kultúraspecifikus szavakhoz köthető: e szavak fordítása igen nehézkes, hiszen sokszor a másik nyelvterület nem is ismeri az adott dolgot a maga valójában, vagy pedig teljesen más képzeteket társít hozzá (*máglyarakás*, *tanyarendszer*). Az egyes típusok különböző altípusokra oszthatók, amelyekre az alábbiakban mutatunk be egy-egy jellegzetes példát:

I. típus: többjelentésű szavak esete

(a) A magyar szónak két angol megfelelője van

Ez akkor okoz problémát, amikor nem lehet a gyakoriságukat vagy fontosságukat rangsorolni, mivel egyenrangúak, egyforma gyakorisággal fordulnak elő, de eltérő

esetekben használjuk őket. Az emberi fordítóknak nem okoz problémát eldönteni, hogy melyik angol szót használják az adott környezetben. Annak érdekében, hogy a gépi rendszer is hasonló megbízhatósággal tudjon dönteni a MorphoLogic fejlesztői egy szemantikai modul integrálásán dolgoznak.

pl:
magyar: *körte*
angol: *light bulb / pear*

(b) A magyar szónak több jelentése van, de a rendelkezésre álló kétnyelvű szótárban a (leg)ritkábban használatos van megadva

Nagyon sokszor találkoztunk olyan szavakkal, amelyeknek több eltérő jelentésű használata van, és igencsak nagy meglepetést okozott a kétnyelvű szótár által adott jelentés. Ilyen pl. a *ráhajtás*, amelynek az angol szótári megfelelője az „*over*" szó (kötésben ráhajtás). A javítást végző csapatban 5-6 egyéb jelentést gyűjtöttünk, melyek között a szótári megoldás nem szerepelt.

(c) Nehéz eldönteni a magyar jelentések esetén a használati sorrendet

Ide azok a jól körülhatárolható, egyértelmű magyar jelentéssel és angol fordítással rendelkező szavak tartoznak, amelyek használatának gyakorisági sorrendje nem állapítható meg egyértelműen, az egyes jelentések szubjektív döntés alapján kerülnek besorolásra. Ide tartoznak pl. a *sitt, szivornya, fazon*, stb.

II. típus: Szótári információk nem helytállóak

(a) Kifordított szótárból fakadó problémák

Az automatikus előfordítás során, részben a MetaMorpho angol-magyar fordítórendszerének szótári adatbázisa került visszafordításra. Emiatt előfordult olyan eset, amikor a bejegyzés magyar fordítása hibás volt, így a visszafordítás során ugyancsak félrevezető eredményt adott. Ilyen például:

angol eredeti: *pharaoh's serpent / pharaoh's rat*
magyar visszafordított: *fáraókígyó / fáraópatkány*
helyes magyar kifejezés: *a fáraó kígyója / egyiptomi mongúz*

(b) A magyar szó jelentése jól körülhatárolható, de hibásan szerepel a szótárban

Ez volt a leggyakrabban előforduló probléma. Ilyenkor hosszabb-rövidebb internetes kutatómunka alapján kerestük meg az alkalmas fordítást. Ilyen volt pl. a *vattapamacs,* amelyhez a szótárban megadott fordítások közül a „*swab*" (fültisztító) jelentése állt legközelebb. Ebben a konkrét esetben pl. angol drogériák internetes katalógusait használtuk. Hasonló példa a *magassági botkormány* is, amikor is egy repülési lexikont böngészve sikerült rábukkanni a megfelelő kifejezésre.

(c) A magyar jelentés jól körülhatárolható, de nem szerepel a szótárban

Általában a speciális szakkifejezések sorolhatók ebbe a kategóriába, mint például a *rázószekrény*. Itt is a jól bevált internetes keresés hozta meg a megfelelő eredményt: kombájnok magyar és angol részletes műszaki leírásait tanulmányozva sikerült megtalálni az angol megfelelőt.

(d) A magyar jelentés nem egyértelműen körülhatárolható és nem szerepel a szótárban, vagy a szótárban szerepel ugyan fordítás, de az egyik magyar jelentéssel sem f ér össze

Ez egy elég nehezen kezelhető csoport, hiszen ha a magyar jelentést sem tudjuk megragadni, akkor a legritkább esetben tudjuk csak megtalálni a helyes angol megfelelőt. Ilyen volt például a *számolócédula*, ami elsőre a mellékszámítások elvégzésére szolgáló kis papírfecninek tűnt. Az internetes keresés viszont az aprónyomtatványok között ismerte fel, amelyből kiderült, hogy a pincérek által használt, reklámgrafikával díszített írótömb is ezen a néven ismert. A szótárban szereplő fordítása, a „*bill slip*" viszont a legjobb esetben is csak a pénztárblokk megfelelője. Ilyen esetekben először a lehetséges magyar jelentéseket gyűjtöttük össze, majd azoknak külön-külön megkerestük a fordítását. Első helyre a leggyakrabban előforduló magyar jelentés megfelelőjét írtuk.

(e). Egyértelmű magyar jelentés, de a kétnyelvű szótárban megadott jelenést nem ismeri az egynyelvű szótár

Az egyik legmeglepőbb példa a *szülőszék* esete, amelyre az EISZ[71] a „*lasanum*" fordítást adja. Ezt azonban nem ismeri egyetlen általunk használt egynyelvű szótár sem. Internetes keresésnél a Google 16 találatot ad rá, amelyek jelentését „*chamber pot*"-ként (bili) határozzák meg. Ez egy különösen sajnálatos eset, hiszen az angol kifejezés is éppoly beszédes, mint magyar megfelelője: „*birthing chair*".

III. típus: Kultúraspecifikus szavak

A kultúraspecifikus szavak esetében sem használható jól a kétnyelvű szótár, legalábbis az első jelentés meghatározásakor. Itt az okozza a problémát, hogy igen nagy eséllyel ezeknek a szavaknak természetüknél fogva nem létezik az angol megfelelője. Ilyenek például a *máglyarakás, sárarany*, vagy éppen az *aranykorona*. Ebben az esetben az ilyenkor szokásos fordítói eljárást alkalmaztuk: a lehető legrövidebb és legpontosabb körülírást adtuk meg hozzájuk.

IV. Egyéb, atipikus problémák

(a) Hangutánzó, hangfestő szavak

Általában ezek a szavak nem, vagy esetleg helytelenül szerepelnek a kétnyelvű szótárakban; pl.: *rotyogás*. Ha a szabály javítója nem ismeri anyanyelvi környezetben szerzett tapasztalatai alapján az ilyen jellegű szavakat, akkor támpont nélkül elég nehezen talál hozzájuk elfogadható megfelelőt. Ilyen esetekben a leírt jelenséghez kapcsolódó körülírásokat kerestük, és azok előfordulási gyakoriságát vizsgáltuk. A „*rotyogás*" például jól körülírható a „*boiling sound*" kifejezéssel, és ennek brit angol használata alkalmasnak és elegendően gyakorinak tűnt.

(b) Alapszókincs megítélése

Sok esetben nehéz eldönteni, hogy mi tartozzon bele az alapszókincsbe. A szavak ismerete, használati gyakorisága az anyanyelvi nyelvhasználóknál is szubjektív, életmód-, lakhely vagy családi háttér-függő. Jól illusztrálja ezt az a példa, hogy az egyik javító számára a *vejsze* szó ismert, sőt általánosan használt volt, míg a *vasgyúró* nem – a többség számára pedig pont fordított volt a helyzet.

(c) Szavak, amelyek fordítása egyik irányban magától értetődő, visszafelé pedig megtévesztő lehet

[71] http://www.eisz.hu/

Akadnak olyan szavak, amelyek "becsapósak" lehetnek a fordító és/vagy a szabályt ellenőrzők számára; ilyen például a magyar *winchester* (angol: *HDD / Winchester*). Míg angolból semmi problémát nem okoz a fordítás, visszafelé nem biztos, hogy annyira egyértelmű.

Értelemszerűen külön zavaró tényező, amikor a fent felsorolt esetek nem tisztán, hanem vegyesen fordulnak elő. Az esetek többségében ez volt a jellemző.

4.2 A melléknévi csoportok fordítása során szerzett tapasztalatok

A melléknévi kifejezések fordítása időigényesebbnek bizonyult, mint a főnévi csoportoké. Ennek oka a szóanyagban kereshető: egyfelől sokkal több az elvont kifejezés (*árválkodó*, *kuláns*, *fellengző*), másfelől igen sok a speciális szakterülethez tartozó szó (*kápolnakoszorús*, *jogdíjköteles*). Ugyanakkor sokkal több szó került kiszűrésre, mivel a nyelvtan képes kezelni a produktív képzéseket (például melléknevek fokozása (*legkisebb*) vagy ható képzős alakok (*bírálható*)), ezért ezeket az alakokat nem vesszük fel a szótárba, hiszen csak feleslegesen növelné a szótár méretét. Kivételt képeznek a rendhagyó alakok, amelyeket természetesen szerepeltettünk a szótárban (pl.: *látható – visible*).

A melléknévi csoportok fordítása során felmerültek bizonyos problémák, amelyek nehézséget jelenthetnek a fordítóprogram számára. Hat csoportba osztottuk a tipikus problémákat. Igyekeztünk mindegyik problémára olyan megoldást találni, ami képes elősegíteni a fordítóprogram fejlesztését, későbbi tökéletesítését.

Az első problémát azok a melléknevek jelentik, amelyek önmagukban sohasem fordulnak elő (*nevű*, *színű*), kötelezően kíséri őket egy másik melléknév vagy számnév: **a nevű fiú* vagy **a színű pulóver* kifejezések agrammatikusak, szemben az *a Feri nevű fiú* vagy *a sárga színű pulóver* kifejezésekkel. Ezek a melléknevek ;HALFLEX megjegyzést kaptak:

```
HU.ADJX = ADJ(lex="színű")
EN.ADJX = ADJ[lex="coloured"]
;HALFLEX
```

A második problémakörbe azok a melléknevek sorolandók, amelyek egyik jelentésükben ;HALFLEX megjegyzést kapnának (l. előző probléma), de más jelentésben szerepelhetnek önállóan is (*éves*): *tizenöt éves háború*, ellenben *az éves jelentés*. Ezeknek a mellékneveknek megadtuk az önállóan használható fordítását, és ;ADJHALFLEX kommenttel láttuk el, így jelölve, hogy más használata is lehetséges az adott szónak:

```
HU.ADJX = ADJ(lex="éves")
EN.ADJX = ADJ[lex="annual"]
;ADJHALFLEX
```

Harmadszor, bizonyos melléknevek angol fordítása függ attól is, hogy a melléknév éppen attributív vagy predikatív pozícióban fordul elő (*kérdéses*). A jelzői funkcióra egy példa: *a kérdéses termék* (*the product in question*), illetve a névszói-igei állítmány névszói részének betöltésére is egy illusztráció: *A játéka még kérdéses.* (*His play is still doubtful.*) Ilyen esetben mindkét jelentést megadtuk, a jelzői jelentést

írtuk fel az angol oldalon, a predikatív jelentést pedig a szabály megjegyzés részében tüntettük fel:

HU.ADJX[enpos=POST] = ADJ(lex="kérdéses")
EN.ADJX(:head=”PREP”) = PREP[lex="in"] + N[lex=”question”]
;PREDIC: doubtful

A negyedik csoportba azok a szavak kerültek, amelyek főnevek és melléknevek is lehetnek, tipikusan ilyenek a népnevek (*angol, holland*), de más típusú szavaknál is előfordult ez a jelenség (*százlábú*, *csuhás*). Melléknévként fordítottuk őket, majd ;ASNOUN kommenttel láttuk el őket, és a főnévi jelentésüket is megadtuk:

HU.ADJX = ADJ(lex="holland")
EN.ADJX = ADJ[lex="Dutch"]
;ASNOUN: Dutchman

Az ötödik esetben a melléknév leggyakoribb használatában olyan szókapcsolatban fordul elő, amelynek angol megfelelője egy szó (*nyolcvanas* (*évek*) vs. *eighties*), ugyanakkor más jelentései is léteznek: *nyolcvanas férfi* (azaz nyolcvan évesnél idősebb, vagy 1980-ban született). Ilyen esetekben a leggyakoribb használatnak megfelelő fordítást adtuk meg, feltüntettük a második és harmadik jelentést is, és ;NUM kommenttel jeleztük, hogy (év)számot tartalmazó kifejezésről van szó:

HU.ADJX = ADJ(lex="nyolcvanas")
EN.ADJX(:head="N") = N[lex="eighty", num=PL]
;sense2: EN.ADJX(:head="PREP") = PREP[lex="in"] + PRON[lex="his"] + N[lex="eighty", num=PL]
;sense3: EN.ADJX = ADJ[lex="born"] + PREP[lex="in"] + NUM[lex="1980”]
;NUM

A hatodik problémakör legfőképpen a leggyakoribb mellékneveket érinti, vagyis hogy a melléknév pontos fordítása igen gyakran az utána következő főnév függvénye (*nagy*, *szép*): *nagy üzlet* (***big*** *business*) vs. *nagy szerep* (***large*** *role*) vs. *nagy siker* (***great*** *success*) vs. *nagy felbontás* (***high*** *resolution*). A példában a melléknévnek legalább négyféle helytálló fordítása is lehet, azonban az adott főnév mellett (többnyire) csak egy adott fordítás a megfelelő. Mivel azonban a szabályban csak a melléknév szerepel, ezekben az esetekben igyekeztünk elsőként a legáltalánosabb jelentést megadni, de a specifikusabb jelentéseket is felsoroltuk:

HU.ADJX = ADJ(lex="nagy")
EN.ADJX = ADJ[lex="big"]
;sense2: EN.ADJX = ADJ[lex="large"]
;sense3: EN.ADJX = ADJ[lex="great"]
;sense4: EN.ADJX = ADJ[lex=”high”]

Vonzatos melléknevek esetében a kötelező vonzatot (ragos vagy névutós főnevet) ;CASE kommenttel adtuk meg. Többjelentésű melléknevek esetében külön gondot jelentett, hogy nem mindegyik jelentésre volt érvényes minden, a szórendre vagy szóhasználatra vonatkozó információ. Ilyenkor ;ADJVAL kommentben megjegyeztük, hogy melyik jelentésre melyik információ vonatkozik.

4.3 A határozószói csoportok fordítása során szerzett tapasztalatok

A határozószói csoportok fordítása bizonyult a legkevésbé problémásnak. Ennek két fő oka volt: egyfelől nagyságrendileg kevesebb határozószó szerepel a szótárban a többi szófajhoz képest, a névszói kifejezések mindössze 4%-a volt ADVX. (Összehasonlításképpen: a névszói csoportok 72%-a főnévi csoport, 24%-a melléknévi csoport volt.) Másfelől pedig a határozószavak – a főnevekkel és melléknevekkel összevetve – igen ritkán többértelműek, így fordításuk is jóval könnyebbnek bizonyult. Egy példa a határozószói szabályokra:

```
HU.ADVX = ADV(lex="baloldalt")
EN.ADVX(:head="PREP") = PREP[lex="on"] + DET[dettype=DEF] +
ADJ[lex="left"] + N[lex="side"]
```

5 Kollokációk

A fordítóprogram minél hatékonyabb működése érdekében a szótári adatbázisba kollokációk is bekerültek. Kollokációnak tekintettünk minden olyan többtagú kifejezést, amelynek tagjai viszonylag gyakran szerepelnek együtt, és formájuk többékevésbé rögzített [4]. Néhány példa: *gyáva nyúl*, *hatos lottó* (NX-k), *gyengén látó*, *kreol bőrű* (ADJX-k), *ízig-vérig* (ADVX) és *eb ura fakó* (idióma). Ezek fordítása a legnehezebb, hiszen a kollokációk nem teljes mértékben kompozicionálisak (vagyis jelentésük nem számítható ki alkotórészeik jelentéséből és azok összekapcsolódási módjából), így a kifejezés részeinek lefordításából előállt szókapcsolat a legtöbb esetben nem tekinthető a kifejezés angol megfelelőjének [8]. Ebből következően az automatikusan generált fordítás igen kevés esetben volt elfogadható, magunknak kellett megtalálni a kollokáció pontos angol megfelelőjét. Nehezítette a munkát az is, hogy igen sok szaknyelvi – különösen jogi és gazdasági – terminus szerepelt a kollokációk között (például *járulékfizetési hajlandóság*, *külkereskedelmi mérleghiány*), amelyek fordítását sokszor még a kétnyelvű szakszótárak sem adták meg.

6 További fejlesztések

Jelenleg a szótárban szereplő főneveket különféle szemantikai jegyekkel látjuk el – többek között *abstract*, *human*, *animate*, *currency*, *bodypart*, *mass* jegyeket használunk –, továbbá feltüntetjük a nyelvtani nemet és a megszámlálhatóságot is. Megfelelő nyelvtani szabályokkal kiegészítve a fordítóprogram így könnyebben tud kezelni bizonyos nyelvtani jelenségeket (például névmási referencia), ezáltal pontosabbá válik a létrejövő fordítás.

A szemantikai jegyek bejelölésének ugyanakkor fontos szerepe van az igei vonzatkeret kitöltésében is: például a *kifizet* ige tárgya csak valamilyen pénznem lehet, azaz kötelezően rendelkezik currency=YES jeggyel:

```
HU.VP = SUBJ + TV(:lex=”kifizet”) + OBJ(pos=N, case=ACC, currency=YES)
Az üzletember kifizetett százezer forintot.
```

Az igei vonzatkeret meghatározása és a főnevek szemantikai jegyeinek megadása a többjelentésű szavak fordítását is megkönnyíti. Például az *aláír* ige mellett a *perjel* szó csak 'egyházi elöljáró' (angolul *prior*) jelentésben fordulhat elő, mivel a *perjel* kétféle jelentése közül csak a *prior* rendelkezik a human=YES értékkel, a *slash* természetesen human=NO értékű.

HU.VP = SUBJ(human=YES) + TV(:lex="alá|ír") + OBJ(pos=N, case=ACC)
*A perjel aláírta az iratokat. – The prior / *slash signalled the documents.*

7 Összegzés

A cikkben összefoglaltuk a MetaMorpho magyar-angol fordítóprogram kétnyelvű szótári adatbázisának előállítása során szerzett tapasztalatokat. Röviden vázoltuk azokat a tipikus problémákat, amelyek a szavak többértelműségéből, illetve a szótárak pontatlanságából adódtak. Említést tettünk a kollokációk fordításával kapcsolatos nehézségekről is. A jelenleg zajló fejlesztések – a szemantikai jegyek bejelölése – a fordítóprogram további tökéletesítését teszik lehetővé.

Bibliográfia

1. Arnold, D. J., Balkan, L., Meijer, S., Humphreys, R. L., Sadler, L.: Machine Translation: An Introductory Guide. Oxford: NCC Blackwell (1994)
2. Isabelle, P.: Electronic Dictionaries and Machine Translation Systems. In: Proceedings of the International Symposium on Electronic Dictionaries (ISED-88). Tokyo (1988)
3. Prószéky, Gábor; László Tihanyi: MetaMorpho: A Pattern-Based Machine Translation System. In: Proceedings of the 24th 'Translating and the Computer' Conference, 19–24. ASLIB, London, United Kingdom (2002.)
4. Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.): Proceedings of CICLING-2002. Mexico City (2002)
5. Tihanyi, L.: A MetaMorpho projekt története. In: Alexin, Z., Csendes, D. (eds.): MSzNy 2003 – I. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem (2003) 247–252
6. Tihanyi, L.: A MetaMorpho projekt 2004-ben. In: Alexin, Z., Csendes, D. (eds.): MSzNy 2004 – II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem (2004) 85–87
7. Tihanyi, L.: A MetaMorpho fordítóprogram projekt 2005-ben. In: Alexin, Z., Csendes, D. (eds.): MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem (2005) 99–107
8. Váradi, T.: Többszavas kifejezések kezelése MT szótárban. In: Alexin, Z., Csendes, D. (eds.): MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem (2005) 233–244